Statistische und stochastische Grundlagen

Vorlesungsmitschrieb zum Modul an der Universität Stuttgart

Simon König

Inhaltsverzeichnis

1

Kapitel 1: Einführung

1: Themenbereiche

1.1 Desktriptive Statistik

1.2 Explorative Statistik

1.3 Induktive Statistik

2: Grundbegriffe

2.1 Grundbegriffe der Statistik

Statistische Einheit Objekte die erfasst werden

Grundgesamtheit Menge aller für die Fragestellung relevanten statistischen Einheiten

Teilgesamtheit Teilmenge der Grundgesamtheit

Stichprobe Tatsächlich untersuchte Teilmenge der Grundgesamtheit

Merkmal, Variable Größe von Interesse

Ausprägung, Wert Konkreter Wert des Merkmals für eine statistische Einheit

2.2 Charakterisierung der Merkmale

diskret abzählbar

stetig Werte aus einem Intervall

quasi-stetig stetig, aber nicht stetig messbar

2.3 Skalen von Merkmalen

Nominalskala Namen, Kategorien

Ordinalskala Ausprägungen mit Ordnung, aber Abstände nicht interpretierbar

Kardinalskala metrisch, messbar

2.4 Datengewinnung, Datenerhebung

Experiment - Erhebung

3: Statistik

3.1 Erste Kenngrößen

Als *Urliste* bezeichnet man die Menge der Merkmale X der Untersuchungseinheiten $U = \{x_1, \ldots, x_n\}$.
Die *auftretenden Ausprägungen* von X sind die Werte $\{a_1, \ldots, x_n\} \subseteq \{x_1, \ldots, x_n\}, k \leq n$. Oftmals treten in einem großen Datensatz der Größe n nicht auch n verschiedene Werte x_i auf.
Damit definieren sich

Definition 3.1: Absolute Häufigkeit

Die absolute Häufigkeit einer auftretenden Ausprägung a in einer Urliste U ist

$$h(a) = |\{x \in U \mid x = a\}|.$$

Es gilt immer, dass die Summe aller absoluten Häufigkeiten gleich der Datensatzgröße ist

$$\sum_{i=1}^{n} h(a_i) = |U|.$$

Die absolute Häufigkeitsverteilung ist dargestellt durch die Folge von Werten

$$h_1, \ldots, h_k = h(a_i), \ldots, h(a_k)$$

Eine grafische Darstellung der absoluten Häufigkeitsverteilung nennt man ein *Histogramm*.

Definition 3.2: Relative Häufigkeit

Die relative Häufigkeit einer auftretenden Ausprägung a in einer Urliste U ist

$$f(a) = \frac{h(a)}{|U|}.$$

Es gilt ähnlich wie bei der absoluten Häufigkeit für die Summe

$$\sum_{i=1}^{n} f(a_i) = 1.$$

3.1.1 Kumulierte Häufigkeiten

Die absolut kumulierte Häufigkeitsverteilung ist die Funktion

$$H(x) = \sum_{i:a_i \leq x} h_i$$

Ebenso, die relative kumulierte Häufigkeitsverteilung

$$F(x) = \sum_{i:a_i \leq x} f_i$$

3.2 Klassifizierung

Sind alle auftretenden Ausprägungen Elemente eines Interfalls $[a, b]$, lässt sich dieses in gleich große Klassen der Größe d unterteilen.
Eine Klassifizierung ist allgemein

$$[a, c_1), \ldots, [c_i, c_{i+1}), [c_{i+1}, c_{i+2}), \ldots \quad \forall i : c_{i+1} - c_i = d$$

Klassifizierte Daten sind i.A. einfacher zu interpretieren als große Mengen von Daten, die sich nur wenig voneinander unterscheiden.
Der maximale Fehler bei der Klassifizierung ist die halbe Klassengröße.

3.3 Lagemaße

Lagemaße helfen beim Vergleich verschiedener Eigenschaften, bzw dem Vergleich verschiedener statistischer Einheiten mit einer gemeinsamen Eigenschaft.

Definition 3.3: Lagemaß

Ein *Lagemaß* ist eine Abbildung $L : \mathbb{R}^n \to \mathbb{R}$ mit der Eigenschaft

$$L(x_1 + a, \ldots, x_n + a) = L(x_1, \ldots, x_n) + a \quad \forall a, x_i \in \mathbb{R} \ (1 \leq i \leq n)$$

Ein Lagemaß beschreibt das Zentrum einer Verteilung.

Beispiele für Lagemaße sind

3.3.1 Arithmetisches Mittel

Das arithmetische Mittel ist nur für quantitative Merkmale sinnvoll. Es berechnet sich durch

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} (a_i \cdot f_i)$$

aus Rohdaten beziehungsweise aus den Häufigkeitsdaten.
Mit dem arithmetischen Mittel gilt die sogenannte Schwerpunkteigenschaft

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0$$

Wie aus der Formel erkennbar, ist das arithmetische Mittel extrem empfindlich gegen Ausreißer. Dafür wurden die folgenden Mittel eingeführt.

Das getrimmte Mittel Um Ausreißer weniger stark ins Gewicht fallen zu lassen wird der Datensatz absichtlich verkleinert. Beim getrimmten Mittel aus einer sortiert vorliegenden Liste von Daten werden zum Beispiel die oberen und unteren 5% der Daten abgeschnitten, damit fallen auch eventuelle Ausreißer raus. Die Datensatzgröße bleibt jedoch nicht erhalten.

Das winsorisierte Mittel Ähnlich wie beim getrimmten Mittel wird der Datensatz beim winsorisierten Mittel von oben und unten herein bearbeitet. Anstatt Daten zu löschen werden beispielsweise die oberen 5% durch den nächstkleineren Wert ersetzt. Hierbei bleibt also die Datensatzgröße gleich.

3.3.2 Median

Der Median stellt ein robusteres Lagemaß als das arithmetische Mittel dar, er ist resistenter gegen Ausreißer im Datensatz. Für $x_1 \leq x_2 \leq \ldots \leq x_n$, also einen sortiert vorliegenden Datensatz ist der Median

$$x_{\text{med}} = \begin{cases} x_{\frac{n+1}{2}} & n \text{ ungerade} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & n \text{ gerade} \end{cases}$$

Benötigt eine Zahlenordnung (oridnal).

3.3.3 Modus